Biometrics 000, 000–000 DOI: 000
000 0000

Web-based Supplementary Materials for "Reinforcement Learning Strategies for Clinical Trials in Non-small Cell Lung Cancer" by Yufan Zhao, Donglin Zeng, Mark A. Sosinski, and Michael R. Kosorok

Yufan Zhao[1], Donglin Zeng[2], Mark A. Socinski[3], and Michael R. Kosorok[2,*]

[1]Global Biostatistics and Epidemiology, Amgen Inc.,

One Amgen Center Drive, Thousand Oaks, California 91320, U.S.A.

[2]Department of Biostatistics, University of North Carolina at Chapel Hill,

3101 McGavran-Greenberg, CB 7420, Chapel Hill, North Carolina 27599, U.S.A.

[3]Department of Medicine, University of North Carolina at Chapel Hill,

Physicians Office Building , 170 Manning Drive, Chapel Hill, North Carolina 27599, U.S.A.

email: kosorok@unc.edu

Web Appendix A: Consistency of Estimating Optimal Treatment Regimes

Using the notation in the paper, we denote π_1 as the decision $\{d_1\}$ at first line and denote π_2 as the decision $\{d_2, T_M\}$ at second line. Moreover, we let $T_1(\pi_1)$ be the potential survival time after first line treatment but before second line treatment. We also let $T_2(\pi_1, \pi_2)$ be the potential survival time after second line treatment given $T_1(\pi_1) = t_2$. Additionally, S_1 denotes the states at first line and $S_2(\pi_1)$ denotes the state at second line. Note that the potential survival time is $T_1(\pi_1) + T_2(\pi_1, \pi_2)I(T_1(\pi_1) = t_2)$. Under a counterfactual framework, to find the optimal treatment strategy, we need to maximize the value function $E_{\pi_1}\left[T_1(\pi_1) + I(T_1(\pi_1) = t_2)\max_{\pi_2} E_{\pi_1,\pi_2}[T_2(\pi_1, \pi_2)|S_2(\pi_1), T_1(\pi_1) = t_2]\Big|S_1\right]$, where E_{π_1} denotes the expectation when the decision at the first line is set to be π_1, and E_{π_1,π_2} means the expectation when the decision at the second line is also set to be π_2. Specifically, the optimal treatment regime can be obtained via the Q-learning algorithm:

$$\pi_2^*(\pi_1) = \text{argmax}_{\pi_2} E_{\pi_1,\pi_2}[T_2(\pi_1, \pi_2)|S_2(\pi_1), T_1(\pi_1) = t_2], \tag{A.1}$$

and then

$$\pi_1^* = \text{argmax}_{\pi_1} E_{\pi_1}\left[T_1(\pi_1) + I(T_1(\pi_1) = t_2)E_{\pi_1,\pi_2^*(\pi_1)}[T_2(\pi_1, \pi_2^*(\pi_1))|S_2(\pi_1), T_1(\pi_1) = t_2]\Big|S_1\right]. \tag{A.2}$$

We now justify that the above functions of the potential outcomes in $(A.1)$ and $(A.2)$ can be estimated consistently via the observed data under our sequential randomized designs. In addition, we assume the consistency assumption (Cole and Frangakis, 2009), i.e., the times T_1 and T_2 satisfy

$$T_1 = \sum_{d_1} T_1(d_1)I(D_1 = d_1), \quad T_2 = \sum_{d_1,d_2,t} T_2(d_1, (d_2, t))I(D_1 = d_1, D_2 = d_2, T_M = t).$$

Following Murphy (2005a), the sequential randomized assumption implies that

$$E_{\pi_1,\pi_2}[T_2(\pi_1, \pi_2)|S_2(\pi_1), T_1(\pi_1) = t_2] = E[T_2(\pi_1, \pi_2)|S_2(\pi_1), T_1(\pi_1) = t_2, D_1 = \pi_1, (D_2, T_M) = \pi_2],$$

so it is equal to $E[T_2|S_2, T_1 = t_2, D_1 = \pi_1, (D_2, T_M) = \pi_2]$ by the consistency assumption.

Similarly,

$$
\begin{aligned}
& E_{\pi_1}\left[T_1(\pi_1) + I(T_1(\pi_1)=t_2) E_{\pi_1,\pi_2^*(\pi_1)}[T_2(\pi_1,\pi_2^*(\pi_1))|S_2(\pi_1), T_1(\pi_1)=t_2]\Big|S_1\right] \\
= \; & E_{\pi_1}\left[T_1(\pi_1) + I(T_1(\pi_1)=t_2) \max_{d_2,t} E[T_2|S_2, T_1=t_2, D_1=\pi_1, D_2=d_2, T_M=t]\Big|S_1\right] \\
= \; & E\left[T_1(\pi_1) + I(T_1(\pi_1)=t_2) \max_{d_2,t} E[T_2|S_2, T_1=t_2, D_1=\pi_1, D_2=d_2, T_M=t]\Big|S_1, D_1=\pi_1\right] \\
= \; & E\left[T_1 + I(T_1=t_2) \max_{d_2,t} E[T_2|S_2, T_1=t_2, D_1=\pi_1, D_2=d_2, T_M=t]\Big|S_1, D_1=\pi_1\right].
\end{aligned}
$$

Therefore, both functions regarding the potential outcomes in the right-hand sides of $(A.1)$ and $(A.2)$ can be estimated via estimating $E[\hat{T}_D|S_1, D_1]$ and $E[T_2|S_2, T_1=t_2, D_1, D_2, T_M]$ using censored data, where $\hat{T}_D$ is as defined in Step 4 of the estimation algorithm given in Section 3. In our paper, we estimate these conditional expectations via ϵ-SVR-C.

Web Appendix B: Example Matlab Code Used in the Simulation

The simple MATLAB code example for the simulation study in Section 4 is available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

References

Cole, S. R. and Frangakis, C. E. (2009). The consistency statement in causal inference. A definition or an assumption? *Epidemiology* **20,** 3–5.

Murphy, S. A. (2005a). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* **24,** 1455–1481.